THE EVOLUTION OF POLYSEMY IN CHILD LANGUAGE

BERNARDINO CASAS[1], NEUS CATALÀ[2], RAMON FERRER-I-CANCHO[2], JAUME BAIXERIES[1]

[1] *Complexity and Quantitative Linguistics Lab. LARCA Research Group. Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya. Campus Nord, Edifici Omega, Jordi Girona Salgado 1–3. Barcelona 08034, Catalonia.*

[2] *Complexity and Quantitative Linguistics Lab. TALP Research Center. Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya. Campus Nord, Edifici Omega, Jordi Girona Salgado 1–3. Barcelona 08034, Catalonia.*

It has been hypothesized that early stages of language have left traces of simpler forms of language, for instance, in child language (Bickerton, 1990; Jackendoff, 1999). Word learning biases in children (Saxton, 2010) suggest constraints that human language had to meet at its very origin. Here a candidate for a new (to the best of our knowledge) learning bias is investigated: a global preference for words with a small number of meanings unraveled by the analysis of massive electronic corpora in English and Dutch.

With the help of the Childes database (MacWhinney, 2000), we have studied the temporal evolution of the polysemy of transcripts in 14 Dutch children and 60 English children. Mothers, fathers and investigators are used as controls. The polysemy of a word is defined by its number of meanings (*synsets*) according to a dictionary: WordNet 3.1 for English (Fellbaum, 1998) and Cornetto 2.0 (Vossen et al., 2013) for Dutch. Thus, the polysemy of a transcript of the speech of an individual is defined as the mean polysemy over all the word tokens that have at least one synset. Only the synsets corresponding to the part-of-speech category of a token are considered. To control for word productivity and transcript length, the same number of word tokens are taken per transcript (if a transcript does not reach that number, it is discarded).

Our analysis shows that the proportion of individuals with significant positive correlations between transcript polysemy and time in children in the total of both languages is much larger than those of any adult: 73% in children versus 6% in mothers, and 0% in fathers and investigators. In English this tendency is much clearer than in Dutch: 32,8% in Dutch versus 81,9% in English for children. In general terms, children at the age of about 20 months exhibit a smaller transcript polysemy than adults and converge later (at about 40 months) to adults values.

However, it should not be concluded that children use a more precise and less polysemous language than adults: it is known that children overextend word

meanings in both production and comprehension (Saxton, 2010).

A bias for words of low polysemy is expected by children's need of clear information word meaning (Harris, Golinkoff, & Hirsh-Pasek, 2011) but our finding is non-trivial because combining (a) that children learn the words that they hear the most (Harris et al., 2011) and (b) the most frequent words tend to be more polysemous (Zipf, 1945; Baayen & Moscoso del Prado Martín, 2005), a bias for high polysemy is predicted. Our bias is about the words that are easier to learn by children from the perspective of adult word polysemy and should not be confused with the principle of mutual exclusivity, which implies that words should have at most one meaning (Markman & Wachtel, 1988), because that principle is about how a child maps words into meanings. The relationship between our bias and that principle should be investigated further.

Acknowledgements

This work was supported by the grant BASMATI (TIN2011-27479-C04-03) from the Spanish Ministry of Science and Innovation.

References

Baayen, H., & Moscoso del Prado Martín, F. (2005). Semantic density and past-tense formation in three Germanic languages. *Language*, *81*, 666–698.

Bickerton, D. (1990). *Language and species.* Chicago University Press.

Fellbaum, C. (Ed.). (1998). *WordNet: an electronic lexical database.* Cambridge, MA: MIT Press.

Harris, J., Golinkoff, R. M., & Hirsh-Pasek, K. (2011). Lessons from the crib for the classroom: how children really learn vocabulary. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (Vol. 3, p. 49-65). NY: Guilford Press.

Jackendoff, R. (1999). Possible stages in the evolution of the language capacity. *Trends in Cognitive Science*, *3*(7), 272-279.

MacWhinney, B. (2000). *The CHILDES project: tools for analyzing talk* (Vol. 2: the database, 3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Markman, E., & Wachtel, G. (1988). Childrens use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.

Saxton, M. (2010). Chapter 6: the developing lexicon: what's in a name? In *Child language. acquisition and development* (p. 133-158). Los Angeles: SAGE.

Vossen, P., Maks, I., Segers, R., Vliet, H. van der, Moens, M.-F., Hofmann, K., Tjong Kim Sang, E., & Rijke, M. de. (2013). Cornetto: a combinatorial lexical semantic database for Dutch. In P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for Dutch* (p. 165-184). Springer.

Zipf, G. K. (1945). The meaning-frequency relationship of words. *Journal of General Psychology*, *33*, 251–256.